We may not cooperate with friendly machines

HUMAN–ROBOT INTERACTION

We may not cooperate with friendly machines

In cooperative games, humans are biased against AI systems even when such systems behave better than our human counterparts. This raises a question: should AI systems ever be allowed to conceal their true nature and lie to us for our own benefit?

Michael Rovatsos

Among humans, white lies and other types of deceptive behaviour are generally tolerated when they are underpinned by good intentions, even if they are considered unethical from (at least some) moral philosophical standpoints. Yet the thought of a machine choosing to deceive us deliberately makes us nervous — even when we know that they are acting in our best interest.

In their new paper[1] published in *Nature Machine Intelligence*, Fatimah Ishowo-Oloko et al. investigate what happens to the efficiency of cooperation when machines reveal their non-human nature to their human partners. Despite advances in human-like robotics[2], we are still far away from this becoming a real concern in robotic systems. However, in the area of chatbots like Siri and Alexa, and other digital assistants that interact with humans in constrained 'virtual' contexts, we are likely to soon be able to create digital assistants that could pass for humans.

Ishowo-Oloko et al. conducted an experiment in which human subjects were asked to play an iterated online game against both human and AI opponents. The authors measured how cooperative humans would behave in both cases, juxtaposing this distinction against another variable — whether the human subjects had been told that their opponent was human or not. The experimental results demonstrate that the propensity of humans to cooperate with opponents who they assume to be bots is lower than that toward human opponents.

Different to previous research in this area, where 'anti-AI prejudice' has been widely documented, the experimental set-up in the paper by Ishowo-Oloko et al. involves a state-of-the-art reinforcement learning algorithm[3] that learns to maximize cooperation against human opponents over time. This creates a more interesting situation, where one might reasonably expect human subjects to detect the bot's ability to learn to be cooperative (while also guarding itself against being exploited), and, as a rational response, to cooperate with such a non-human 'benevolently rational' opponent.

The authors demonstrate that even in settings where humans and their AI counterparts can learn from their past interactions, humans do not recover from their initial biases against the non-human opponents, even though a more efficient overall behaviour could be achieved if the bots identified as human. More specifically, they show that the willingness of humans to cooperate early in the game, as well as the ability of bots to achieve high cooperation throughout the game with a human player, are compromised if the human player knows the bot's identity. By contrast, bots are able to achieve and maintain high levels of cooperation (higher than human associates) if the human player believes that they are playing against another human.

Ishowo-Oloko et al. view their experimental results as evidence for what they call a transparency–efficiency tradeoff, suggesting that there may be cases in which being transparent about the true nature of the system might be detrimental to achieving optimal benefit.

The study has its limitations. One of these is that the experimental scenario is based on the academically motivated, abstract game of the iterated prisoner's dilemma[4], which provides a mathematical representation of a small (but hard) moral dilemma. The different strategies and complex dynamics in this game are well-understood by economists, mathematicians and computer scientists, but it is unclear whether the participants understood the complexities of the scenario, and were capable to interpret and reflect on the observed opponent behaviour when making their own decisions.

A related point is that taking transparency seriously would have implied not only informing human participants about the AI nature of their opponents, but also about their learning capabilities, their ability to cooperate when their human opponent does and general benevolence — that is, that they are not trying to trick or exploit the human user.

There are of course a number of ethical issues associated with the use of deceit in AI systems, a topic that was beyond the scope of the paper. The strength of truthfulness and honesty as a normative value lies in the fact that there is only one truth (in factual matters), whereas there are many choices as to how we may deceive. The information provided to a person in each of these choices may substantially (though often subtly) influence their behaviour and restrict their autonomy.

Respect for human dignity also implies that we allow people to act against their own (or anybody else's) best interest. While societies and states introduce rules for rewarding or punishing behaviours, they generally avoid taking individual action choice per se away from their citizens.

On a more practical level, transparency mitigates against flawed definitions of norms and allows for their correction over time through individual resistance and disobedience. If humans do not have all of the information needed to make choices in front of them, how could the users of such systems monitor and correct, for example, the efficiency criteria embedded by designers into AI systems?

We are already observing massive levels of public distrust in the opaque optimization criteria employed by online platforms using AI systems — for example, commercial search engines and recommendation marketplaces. Trusting that they would act in their users' best interest by hiding crucial information about the AI algorithms they use would seem highly imprudent.

That said, useful new ideas for how to use variable levels of transparency may come out of this type of research. The article by Ishowo-Oloko et al. raises thought-provoking questions around the ethical use of transparency in AI systems, and provides an exploratory study into a

complex phenomenon, which will hopefully lead to further investigation. In the context of current controversies around the biases of AI systems towards different types of humans, it would be interesting to see how intelligent algorithms could be used to address such biases through appropriate interventions, including those that stem from human-to-human bias embedded in the data used to train many real-world algorithms. ❐

Michael Rovatsos
School of Informatics, University of Edinburgh, Edinburgh, UK.
e-mail: michael.rovatsos@ed.ac.uk

Published online: 12 November 2019
https://doi.org/10.1038/s42256-019-0117-1

References

1. Ishowo-Oloko, F. et al. *Nat. Mach. Intell.* https://doi.org/10.1038/s42256-019-0113-5 (2019).
2. Chikaraishi, T., Yoshikawa, Y., Ogawa, K., Hirata, O. & Ishiguro, H. *Future Internet* **9**, 75 (2017).
3. Crandall, J. W. et al. *Nat. Commun.* **9**, 233 (2018).
4. Axelrod, R. & Hamilton, W. D. *Science* **211**, 1390–1396 (1981).